Identification of Rhetorical Structure Relation from Discourse Marker in Bengali Language Understanding

Abhishek Sarkar, Pinaki Sankar Chatterjee
School of Computer Engineering
KIIT University, Bhubaneswar, Odisha
{abhishek690 , pinaki.sankar.chatterjee}@gmail.com

Abstract: **Rhetorical Structure Theory (RST) has been in constant research for a number of years. The different parts of a text are called text spans. RST helps in organizing these text spans. These text spans are connected with each other by Discourse markers. In this paper, we are exploring the possibility of realizing some RST relations (CONTRAST, SEQUENCE and PARALLEL) from multi-nuclear sentences in Bengali Language with the help of the semantic structure of the sentences and the discourse markers. We present a rule based approach for comprehending the RST relations from the discourse markers in Bengali Language.**

Keywords **– RST, Discourse Marker, Compound Sentence, Syntactic Aggregation, Semantic Representation.**

I. Introduction

Rhetorical Structure Theory is a theory of organization of the text spans, the working of those text spans and how those text spans are related and connected to each other [2]. Each text span can have one or two roles in a relation: it can be a nucleus or a satellite. Nucleus represents the central unit in the sentence, they exist independently. They are considered essential to the understanding of the text. Satellites is less central and can never exist independently, they are actually dependent on the nucleus. Satellites contribute additional information to the nucleus [3].

Discourse marker is a word or phrase that helps in logically connecting two text spans which are coherent to each other [2]. Examples of Discourse Markers are:- but, then, and, although, yet, with, etc

In this paper, we are dealing with only compound sentences, i.e. sentences having more than one nucleus. Compound sentences are those sentences which are formed by combining two or more simple sentences having one nucleus. Some example are given as:- **(1) CONTRAST:-** Ram is very happy with his marks but his sister is very upset with her marks. Here the discourse marker is "but" and the RST relation is CONTRAST.

(2) SEQUENCE:- Bumba is going to the shop then he will go to visit his grandmother.

Here the discourse marker is "then" and the RST relation is SEQUENCE.

(3) PARALLEL:- Dipshika is eating while watching the television.

Here the discourse marker is "while" and the RST relation is PARALLEL.

There are many more like EXPRESSION, CONJUNCTION, CONTINUATION, LIST, etc.

A. Objective

The work in this paper has been done in Bengali Language. Since the formation of Bengali Language is free order, so Ambiguity occurs in mapping of Discourse markers to the RST relations. Some Discourse Marker can be mapped to more than one RST relation hence there exist some ambiguity in mapping function. So to resolve this ambiguity we have tried to formulate some algorithms for each of the RST relations. For a given Semantic representation of compound sentence, we can identify the RST relation of a sentence with utmost accuracy.

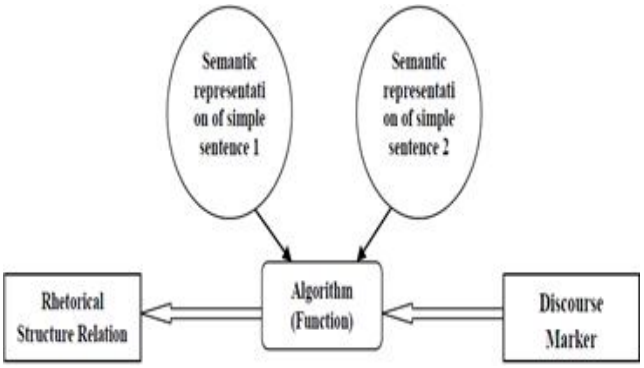

Figure 1: Inputs of the function

The Figure 1 explains the system we are building. Our formulated algorithm or function is going to receive the semantic representations of simple sentence 1 and simple sentence 2 along with the discourse marker and is going to give the corresponding RST relation as the output. So our objective is to identify the RST relation from discourse marker in Bengali Language.

II. Related Works

One of the well known tasks in Discourse marker and RST has been done by *Maite Taboada(***Discourse markers as signals (or not) of rhetorical relations) [1]**,where he discusses the idea of discourse markers or connectives signaling in the presence of a particular relationship. He addressed the relationship between discourse markers and rhetorical relations, and, more generally, the signaling of rhetorical relations. He along with William C. Mann (**Rhetorical Structure Theory: Looking Back and Moving Ahead) [5]** reviewed some of the discussions about RST, especially addressing issues of the reliability of analyses and psychological validity, together with a discussion of the nature

DOI: 03.LSCS.2013.3.53

of text relations.

William C. Mann and *Sandra A. Thompson* (**Rhetorical Structure Theory: Towards a functional theory of text organization) [3]** extended the idea of what actually Rhetorical Structure Theory is. This paper marks the origin of RST. It describes RST as descriptive theory of a major aspect of organization of natural text. This paper establishes a new definitional foundation for RST.

*Daniel Marcu and Abdessamad Echihabi***(An Unsupervised Approach to Recognizing Discourse Relations) [9]** present an unsupervised approach to recognize discourse relations of contrast, explanation-evidence, condition and elaboration that hold between arbitrary spans of texts.

Sumit Das, Anupam Basu, Sudeshna Sarkar (**Discourse Marker Generation and Syntactic Aggregation in Bengali Text Generation) [2],** proposes the idea about the prevalent syntactic aggregation constructs in Bengali. The paper presented a rule based approach towards generating Bengali compound sentences using the identified constructs.

III. The Proposed System

We are only considering compound sentences. There are many types of multi-nuclear RST relations, but in this paper we are only dealing with CONTRAST, SEQUENCE and PARALLEL. In Bengali Language some of the Discourse markers are: - আর, যদিও, কিন্তু, comma (,), তবুও, etc.

We have collected many Bengali compound sentences and divided them manually into 2 different simple sentences. Suppose if we take a Bengali sentence like:- "রাজবাড়িতে সুওরানীর বড় আদর আর দুওরানীর বড় অনাদর।" [10]- This compound sentence can be divided into 2 simple sentences like:- Simple sentence1- "রাজবাড়িতে সুওরানীর বড় অনাদর।" Simple sentence 2 - "রাজবাড়িতে দুওরানীর বড় অনাদর।" The discourse marker is "আর"

Figure 2 shows the semantic representation of the compound sentence[3], where the clause count is given as 2, the type of the sentence is represented as composite, the discourse marker is identified as "আর" and rhetoric relation is given as question mark implying the objective of our paper, which is to identify the RST relation. We have designed algorithms for RST relations CONTRAST, SEQUENCE and PARALLEL.

- The algorithm for CONTRAST:-

```
Algorithm CONTRAST(simple_sentence1, simple_sentence2,
discourse_marker)
{
If(discourse_marker=="কিন্তু"||discourse_marker=="হলেও"||discours
e_marker=="তবুও"||discourse_marker == "তবু" ) then
RST = CONTRAST
Elseif(discourse_marker=="আর||discourse_marker==","||discourse_
marker=="এবং" || discourse_marker == "যদিও" ) then
Begin
if(polarity(simple_sentence1)!=polarity(simple_sentence2)) then
RST = CONTRAST
Else if ((polarity(simple_sentence1) == polarity
(simple_sentence2)) &&((verb (simple_sentence2 ) == antonym
(verb(simple_sentence1))))) then
RST = CONTRAST
End if
Return RST
}
```

```
clause_count: 2
type: composite
rhetoric_relation:?
discourse_marker: আর
sentence_begin:#
    clause_begin:#
        predicate_begin:#
            verb=আদর
            theme=love
            tense=present
            aspect=simple
            polarity=positive
        predicate_end:#
        argument_begin:#
            arg_name=kar
            n-root=সুওরানী
        argument_end:#
        argument_begin:#
            arg_name=kothai
            n-root=রাজবাড়ি
argument_end:#
    clause_end:#
sentence_end:#
sentence_begin:#
    clause_begin:#
        predicate_begin:#
            verb=অনাদর
            theme=hate
            tense=present
            aspect=simple
            polarity=negative
        predicate_end:#
        argument_begin:#
            arg_name=kar
            n-root=দুওরানী
        argument_end:#
        argument_begin:#
            arg_name=kothai
            n-root=রাজবাড়ি
        argument_end:#
    clause_end:#
sentence_end:#
```

Figure 2: After Syntactic Aggregation semantic representation of compound sentence[2] [4].

DOI: 03.LSCS.2013.3.53

Let us take some examples for Contrast:-

Example 1:- "দুওরানীর ঘরে রাজা একটি দিন আসেন কিন্তু সুওরানীর ঘরে রাজা বারো-মাস থাকেন।" "[10]- Here the discourse marker is "কিন্তু" and according to our algorithm, if the discourse marker is কিন্তু, RST is Contrast.

Example 2:- "ছোটরানীর কথা রাজা সবসময় আসেন কিন্তু বড়রানী কথা একবারও ভাবেন না।" [10]- Here the discourse marker is "আর", then we need to check the semantic structure of the simple sentences, here the polarities of the two sentences are different, then RST is Contrast.

Example 3:- "রাজবাড়িতে সুওরানীর বড় আদর আর দুওরানীর বড অনাদর।" [10]–Here the discourse marker is "আর" then we check the semantic structure of the simple sentences, here the polarities of the two sentences are same but the verb "অনাদর" in the simple sentence 2 is an antonym of the verb "আদর" in the simple sentence 1, according to our algorithm RST is Contrast.

- Algorithm for Sequence:-

```
Algorithm SEQUENCE( simple_sentence1 ,simple_sentence2, discourse_marker)
{
If ((discourse_marker == "যখন-তখন") && (polarity (simple_sentence1) == polarity (simple_sentence2))) then
RST = SEQUENCE

Else If (((tense (simple_sentence1) == Past tense) && (tense (simple_sentence2) == Present tense)) || ((tense (simple_sentence1) == Present tense) && (tense (simple_sentence2) == Future tense))) then
RST = SEQUENCE

P1 = PRIORITY (Kakhana)
P2 = PRIORITY (Kakhana)
If (P1 < P2) then
RST = SEQUENCE

If (( verb (simple_sentence1) == non-finite form of a verb) && ( verb (simple_sentence1) == non-repetitive form of a verb)) then
RST = SEQUENCE
Return RST
}
```

- Algorithm for function PRIORITY() :-

```
Algorithm PRIORITY ( Kakhana )
{
If (Kakhana == "ভোরবেলা") then Kakhana_priority = 0
Else if (Kakhana == "সকালবেলা") then Kakhana_priority = 1
Else if (Kakhana == "দুপুরবেলা") then Kakhana_priority = 2
Else if (Kakhana == "বিকেলবেলা") then Kakhana_priority = 3
Else if (Kakhana == "সন্ধ্যাবেলা") then Kakhana_priority = 4
Else if (Kakhana == "রাতেরবেলা") then Kakhana_priority = 5 Else if (Kakhana == "মাঝরাতে") then Kakhana_priority = 6
Else if (Kakhana == "গভিররাত") then Kakhana_priority = 7
Return Kakhana_priority
}
```

Let us take some example for Sequence :- **Example 1:-** "যখন আমি রানী ছিলুম তখন রাজার জন্যে মালা গেঁথেছিলুম।" [10] – Here we can see that the discourse marker is "যখন-তখন" and we can see the polarity of the t wo simple sentences are same. So, according to our algorithm RST is Sequence.

Example 2:- "আজ রাতে রাজা ছোটরানীর ঘরে থাকবেন, কাল হয়তো বড়রানীর ঘরে থাকবেন।" [10] - Here the two simple sentences are in different tense. So according to our algorithm RST is Sequence.

Example 3:- "শঙ্কর সকালবেলা বন্ধুবান্ধবদের বাড়িতে গিয়ে আড্ডা দেয়,বিকেলবেলায় পালঘাটের বাঁওড়ে মাছ ধরতে যায়।" [11] - Here the two simple sentences implies about different time in a single day. We have defined a function PRIORITY( ) where we have divided a single day into 8 different spans and given each of them different priority values. Since the priority value of সকালবেলা is less than বিকেলবেলায় so RST is Sequence.

Example 4:- "রানী গলার গজমতি হার দেখিয়ে বললেন -দেখ রাজা, এ মুক্তো বড় ছোট।" [10] – Here the verb দেখালেন gets changed to দেখিয়ে when the two simple sentences are joined to get the composite sentence and thus gives us a non-finite form of a verb and it's non-repetitive and thus RST is Sequence.

- Algorithm for PARALLEL RST:-

```
Algorithm PARALLEL( simple_sentence1, simple_sentence2, discourse_marker)
{
If ((simple_sentence1(Ke) == simple_sentence2(Ke)) && (tense (simple_sentence1) == tense (simple_sentence2))) then
RST = PARALLEL

If ( verb (simple_sentence1) == non-finite form of a verb) && ( verb (simple_sentence1) == repetitive form of a verb)) then
RST = PARALLEL
Return RST
}
```

Let us take some example for Parallel:-

Example 1:- "ভাঙা ঘরে দুওরানী নীল সাগরের পানে চেয়ে, ছেঁড়া কাঁথায় পড়ে রইলেন।" [10] – Here the actor ("Ke") in both the sentences are same and both the sentences are in same tense. So RST is Parallel.

Example 2:- "বানর নাচতে নাচতে-ভাঙা ঘরে গেল।" [10]– Here the verb "নাচলো" is the continuous form of the verb "নাচ" and gets changed into a repetitive form in the composite sentence "নাচতে নাচতে" and also it is non-finite form of the verb. So RST is Parallel.

IV. Result and Analysis

We have collected 1000 compound or composite sentences from short stories "KHIRER PUTUL" and "CHADER PAHAR", out of these 1000 sentences, we have randomly

DOI: 03.LSCS.2013.3.53

taken 600 sentences as our training data set and the rest 400 sentences as our testing data set.

With the training data set, we have manually categorized sentences into Contrast, Sequence and Parallel depending on the definition of these RST relations, then we have split the composite sentences into 2 simple sentences manually and then create the semantic structure of each of the sentences, after which we manually identify the Discourse marker from each of the sentences. Then we categorized the sentences into Contrast, Sequence and Parallel. we have developed some rules for each of the RST relation from their semantic representation. These rules were then formulated into an algorithm for each of the RST relations. Then we have created a system with the help of our algorithm.

The testing data set i.e. the remaining 400 sentences are first manually categorized into Contrast, Sequence and Parallel depending on the definition of these RST relations and then to check the accuracy of our system, we put the semantic representation of these 400 sentences into our system.

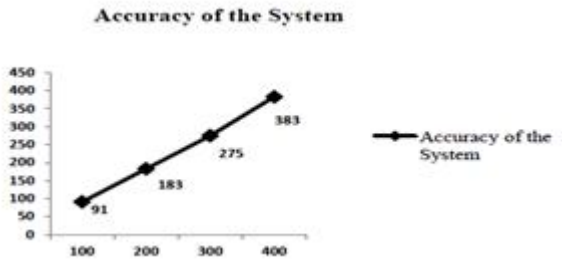

Figure.3:- Line Chart to show the accuracy of our system

The figure 3 shows the accuracy of the system. The accuracy is measured on the testing data set. The horizontal axis represents the Testing Data set, i.e. the number of sentences tested and the vertical axis represents the number of correct identification of RST relation by our system. We can see that for the first 100 sentences, 91 sentences were identified correctly by our system and so on. When all the 400 sentences are tested, 383 RST Relations are correctly identified by our system.

Conclusion

In this paper we have discussed about the identification of the Rhetorical Structure Theory Relation from Discourse Marker in Bengali Language Understanding. We have derived algorithms for CONTRAST, SEQUENCE and PARALLEL RST relation and explained them with few examples. We have given user based evaluation to validate our approach. We have built a small system and a small corpus to check our result.

Future Works

There are many more RST relations whose algorithms can be derived and then a system can be built using the functions of all the RST relations.

References

[1] Maite Taboada : "Discourse markers as signals (or not) of rhetorical relations" in Journal of Pragmatics 38 (2006) 567–592

[2] Sumit Das, Anupam Basu, Sudeshna Sarkar: "Discourse Marker Generation and Syntactic Aggregation in Bengali Text Generation" in Proceedings of the 2010 IEEE Students' Technology Symposium, 3-4 April 2010, IIT Kharagpur

[3] William C. Mann and Sandra A. Thompson: "Rhetorical structure theory: Toward a functional theory of text organization". In Text, 8(3):243- 281, 1988.

[4] S amit Bhattacharya, Sanyog: "An iconic system for multilingual communication for people with speech and motor impairments". M.S. Thesis, IIT, Kharagpur, Supervisor - Anupam Basu and Sudeshna Sarkar, 2004.

[5] Maite Taboada and William C. Mann: "Rhetorical Structure Theory: Looking Back and Moving Ahead" in Discourse Studies 8 (3) by Sage Publicat, January 24, 2006.

[6] Caroline Sporleder, Alex Lascarides:"Using Automatically Labelled Examples to Classify Rhetorical Relations: An Assessment", Printed in the United Kingdom Cambridge University Press, Received 26 August 2005; revised 16 March 2006.

[7] Farhi Marir and Kame1 Haouam: "Rhetorical Structure Theory for content-based indexing and retrieval of Web documents" in Proc. of 2nd International Conference on Information Technology: Research and Education, 2004, ITRE 2004. Pages:160 - 164

[8] Maite Taboada and Manfred Stede: "Introduction to RST Rhetorical Structure Theory" created as part of a project on Natural Language Generation at the Information Sciences, May 2009.

[9] Daniel Marcu and Abdessamad Echihabi: "An Unsupervised Approach to Recognizing Discourse Relations" in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 368-375.

[10] Abanindranath Thakur: "Khirer Putul" Published by Ananda Publications, 1896. (For Corpus use)

[11] Bibhutibhushon Bondopoddhay: "Chander Pahar" Published by Bornayon, 1937 (For Corpus use)

DOI: 03.LSCS.2013.3.53